Systems genetics of complex diseases using RNA-sequencing methods

Mazzoni, Gianluca; Kogelman, Lisette; Suravajhala, Prashanth; Kadarmideen, Haja

Published in:
International Journal of Bioscience, Biochemistry and Bioinformatics

DOI:
10.17706/ijbbb.2015.5.4.264-279

Publication date:
2015

Document version
Publisher's PDF, also known as Version of record

Systems Genetics of Complex Diseases Using RNA-Sequencing Methods

Gianluca Mazzoni, Lisette J. A. Kogelman, Prashanth Suravajhala, Haja N. Kadarmideen*
Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, 1870 Frederiksberg C, Denmark.

* Corresponding author. Tel: +45 35333577; email: hajak@sund.ku.dk
Manuscript submitted: February 25, 2015; accepted April 19, 2015.
doi: 10.17706/ijbbb.2015.5.4.264-279

Abstract: Next generation sequencing technologies have enabled the generation of huge quantities of biological data, and nowadays extensive datasets at different 'omics levels have been generated. Systems genetics is a powerful approach that allows to integrate different 'omics level and understand the biological mechanisms behind complex diseases or traits. In the recent, transcriptomic studies with microarrays have been replaced with the new powerful RNA-seq technologies. This has led to detection of novel gene transcripts, novel regulatory mechanisms, allele specific gene expression and numerous non-coding RNAs (ncRNAs). The integration of transcriptomics data with genomic data in a systems genetics context represents a valuable possibility to go deep into the causal and regulatory mechanisms that generate complex traits and diseases. However RNA-Seq data have to be treated carefully and the choice of the right methodology could have a great impact on the final results. Furthermore the integration of different level is not trivial. Here we give a comprehensive systems genetics overview of the methods and tools for analysis and the integration of RNA-Seq data including ncRNAs. We focused principally on merits and demerits of tools for post mapping quality control, normalization, differential expression analysis, gene network analysis, and integration of different omics data in order to generate a comprehensive guideline to systems genetics analysis using RNA-Seq data.

Key words: Quality control, differential expression, network construction, non-coding RNA, data integration.

1. Introduction

The term "Systems Genetics", a branch of systems biology was originally proposed by Kadarmideen *et al*., which integrate 'omics scale measurements from genome to metabolome to functome through transcriptome and proteome [1]. It assimilates a holistic analysis model to find important causal and regulatory genes and their variants in predicting biomarkers. In recent times, based on high leverage of bioinformatics data that are produced, systems genetics have provided systems level understanding the biological phenomena [2]. This systems genetics approaches have been applied in livestock [3], [4], humans [5] and thoroughly reviewed [6]-[8]. However, most of these previous studies are based on chip-based or array based high throughput 'omics data. With next generation sequencing (NGS) data providing an unprecedented means to construct comprehensive maps of genetic/gene expression variation, there is more to understanding the 'omics approach. This includes several dozens of million "reads" to map single nucleotide variants (SNVs), hundreds of thousands of small insertions or deletions, structural variants and transcripts, epigenetic analysis using ChIP-seq technologies [6], [9].

As more and more technologies enabling huge datasets are available, there is a need for significant understanding and improvement of standard resources and sequencing tools to analyse complex diseases. As a most recent improvement, several non-codingRNA (ncRNA) sequences, *viz.* micro-RNAs (miRNA), long non coding RNA (lncRNA) and recently discovered long intergenic ncRNAs (lincRNAs), competing endogenous RNAs (ceRNA) and enhancer RNAs (eRNA) etc. form a part of regulatory networks and pathways. While it is known that the regulatory networks utilize protein-protein interaction (PPI) data, they are known to affect the expression of protein-coding genes [10]. The ncRNAs have different modes of action and so are classified based on their size and modulation. The smaller among the ncRNAs, miRNAs serve as important modulators of development wherein they regulate transcription by interacting with promoter regions or change protein levels during post-transcriptional stage [11]. This has enabled researchers to focus on diverse transcribed genes between small RNAs and their longer transcripts. However, the association between the latter two categories is beginning to be understood thus allowing researchers to focus more on lncRNAs. Subsequently, a few "meta-analysis" based approaches integrating omics data at the systems genetics level have come up [12]. The approaches integrate several types of 'omics data that include gene expression data, ChIP-seq and miRNA expression data. When these approaches when cross-validated with each other, it will help find the distinct datasets further allowing researchers to comprehensively use these data at the systems genetics level.

Given the paradigm shift of omics data that is processed based on NGS technologies, the challenges of guiding systems genetics/biology research are two-fold, *viz.* analyzing array-based omics datasets and comprehensive understanding of non-coding sequences centered on RNA-Seq data, for example, in the form of long noncoding RNA (lncRNAs). One of the main objectives of this perspective article is to provide an overview of existing tools for analyses based on NGS based RNA-Seq transcriptomics and genome variation datasets and for building up gene networks underlying complex diseases and traits. We focus on the merits and demerits of different RNAseq data analyses approaches. This includes post-mapping quality control and normalization methods and subsequent use of both normalized RNA-Seq data (as normalized reads or counts), differential expression (DE) and network analyses (Fig. 1). We also reason how ncRNA data could be used to integrate 'omics datasets in finding genetic variation and intermediate molecular phenotypes.

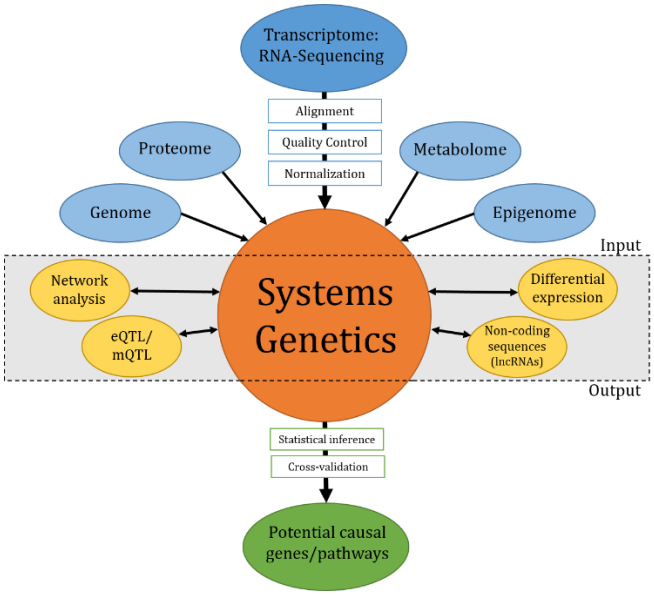

Fig. 1. Overview of systems genetics analysis using RNA-sequencing data.

2. Processing of RNA-Sequencing Data

2.1. Post-mapping Quality Control

After generating RNA-Seq data and alignment (to the reference genome or *de novo* assembly), it is essential to perform a quality control of the aligned reads before proceeding with subsequent analysis of the transcriptome [13]. The quality control allows us to find out any potential bias in the data introduced during previous phases of the workflow: sample extraction, library preparation, sequencing technology, and/or mapping algorithms. Nowadays, there are a lot of freely available tools for the post-mapping quality control. Obviously, these tools do not have a direct impact on the final outcome of the analysis and therefore, there is not a tool that in general performs better than others. Each tool computes different statistics and presents different output types, so the choice of the user is strictly dependent on the type of information and the output format that the user requires. We present here six freely available tools that are widely used: FastQC, Picard Tools, Qualimap, RNASeqQC, RSeQC and SAMStat. In particular, we report the characteristics that are in our opinion most significant for the RNA-Seq data analysis and useful for the users (Table 1). Due to the heterogeneity of the statistics that these tools compute, we grouped them into: BAM/SAM general statistics (e.g. number of reads mapped, mismatches, nucleotide composition, GC-percentage, k-mer bias), RNA specific statistics (e.g. distribution of the reads, genome coverage, intron/exons coverage, intron/exons junctions analysis), mapping quality statistics, insert size statistics (only for paired-end reads), multi-sample comparisons.

Table 1. Overview of Six Widely Used Post-mapping Quality Control Tools, on Alphabetical Order

Tool Name	Enviromen t	Input type*	Output	BAM file QC	RNA-seq specific	Mapping quality	Insert Size	Multi-sample
FastQC [14]	Java	CL & GI	Html page	✓✓	×	×	×	×
Picard Tools [15]	Java	CL	Text file	✓✓	✓✓	×	✓	×
Qualimap [16]	Java and R	CL & GI	Html page	✓	✓✓	✓✓	✓	✓
RNASeqQC [17]	Java	CL	Html page	✓	✓	×	✓	✓
RSeQC[18]	Python	CL	Shell output Text files	✓	✓✓	✓	✓✓	×
SAMStat [19]	C (Unix)	CL	Html page	✓✓	×	✓✓✓	×	×

*GI= Graphical Interface; CL = Command Line.

Picard tools and SAMStat focus mostly on biases related to the library preparation and sequencing process. In particular, SAMStat focuses on the number of mismatches, nucleotide composition and percentage of mapped reads. SAMStat is able to compute the statistics based on the mapping quality to contrast properties of unmapped, poorly mapped and accurately mapped reads to find out whether any properties of the reads influence the mapping accuracy. Picard tools provides a lot of information related to the BAM statistics with a module that is specific for RNA-Seq experiments (CollectRnaSeqMetrics); however, the results are not directly interpretable as output is given as numeric information in a text file. Another important parameter to evaluate is the insert size to detect potential problems during the library preparation or the alignment. While computation of the insert size is implemented in Picard tool, Qualimap, RNASeqQC and RSeQC, the first three extract it directly from SAM file and on the other hand, RSeQC performs a more complex computation. RSeQC takes the eventual presence of introns between two paired reads into consideration, because two paired-end reads that fall in two different exons will potentially have a higher insert size than expected. Among the six evaluated tools, Qualimap and RNASeqQC seem to be the

most comprehensive tools. Both compute statistics for the sample comparison and in particular, RNASeqQCs performs a very useful multisample comparison providing output tables with correlations among samples and GC content. Qualimap is a very user-friendly tool that can be run using both command and graphical line interfaces. With a single command, it automatically gives a html output with easily interpretable tables and graphs. Furthermore, it allows to test the adequacy of the sequencing depth. On the contrary, even though RSeQC computes very usefully and accurately described statistics, the user has to run different modules as some of them do not render the output graphs, but executing the R-scripts separately would solve the purpose. In conclusion, each tool has its own characteristics and featured qualities and the best way is to ascertain the quality of the data using more than one tool. Recently Kroll *et al*. [20] implemented a new tool called Quality Control for RNA-Seq (QuaCRS), which runs a proportion of three different QC tools (FastQC, RNASeQC and RSeQC) and merges the results in a coherent way. In addition, it generates a database that stores and displays the quality control results in an easily accessible way.

2.2. Normalization of RNA-Seq Expression Levels

Normalizing RNA-Seq data is a vital step before analyzing the expression data as it can have a serious impact on the outcome of subsequent analyses. The estimated expression levels of each biological entity need to be normalized in order to compute an accurate comparison between and within samples. There are different types of normalization to account for biases of different nature that could be within or between samples. The within-sample bias includes, e.g., transcript length bias [21], whereas between-sample bias includes the differences in the sequencing depth (library size) and GC-content [22]. This ensures that an accurate type of normalization is dependent on the objective of the study. If the research objective is to detect DE genes, the between-sample bias should be considered. On the contrary, if the research objective is to rank or compare genes within a sample, more complex normalization methods are needed to consider potential bias within samples.

To detect DE genes, the most influencing bias is in general the library size. There are different methodologies that remove this sequencing depth bias. The basic methods consider simply the total count of reads in a sample, but they do not take in consideration the differences in the distribution of the data and the possible incidence of few very highly expressed genes that represent the utmost part of the total reads count. To overcome this issue, researchers have developed methods using scaling factors that match distributions of the samples [22].

Table 2. Overview of Different RNA-Seq Normalization Methods, on Alphabetical Order

Method	Description of the method
DESeq Scaling Factor[23]	The scaling factor for each sample is computed as the median of the ratio between the count of each gene and their respective geometric mean computed across samples
Median	Gene counts are divided by the median of the gene counts
Quantile[24], [25]	Normalize the distribution of counts across lanes
RPKM [26]	Gene counts are divided by the transcripts length times number of millions mapped reads
SAMseq method [27]	Computes the mean read count over the features that result to be null in the dataset.
TC	Gene counts are divided by sequencing depth and multiplied by the average of the total counts
TMM[28]	Genes that are most expressed and with the highest log ratios are removed, and using the remaining genes a scaling factor is computed as the weighted mean of log ratios between the sample and the reference
Upper quartile [29]	Gene counts are divided by the upper quartile of the gene counts

Dillies *et al.* [30] compared a number of normalization methods (Table 2) using a real and a simulated dataset. The best performances were obtained from the Upper Quantile, Median, DESeq, and TMM normalizion, but only DESeq and TMM showed a good precision and sensibility in terms of false positive rate and power of detection. RPKM and TC indicated to be ineffective and so not suggested for DE analysis,

while the Quantile normalization only gave positive results when the samples had equal distributions. On the contrary, in a study from Seyednasrollah *et al.* [31] different DE-tools were tested (baySeq, Cuffdiff 2, DESeq, EBSeq, edgeR, limma, NOIseq, and SAMseq), but they did not find any significant differences between default normalization methods applied by those tools and the TMM normalization. This could because of the probable performance of the normalization phase which is strictly dependent on the characteristics of the dataset.

Table 3. Overview of Methods for Differential Expression Analysis, on Alphabetical Order

Tool	Normalization	Distribution assumptions	Model based statistic	Multifactor	Absence of replicates	Isoforms detection
baySeq [32]	Quantile (TMM, total)	NB	Empirical Bayes	×	×	×
Cuffdiff2[33]	DESeq like (quartile, fpkm)	Beta NB	t-test	×	✓	✓
DESeq [23]	DESeq Scaling Factor	NB*	Exact test	✓	✓	×
DESeq2[34]	DESeq Scaling Factor	NB	GLM	✓	✓	×
edgeR[35]	TMM (upper quantile, DESeq like)	NB	Exact test	✓	✓	×
Limma [36]	TMM	Voom transformation	Empirical Bayes	✓	×	×
NOISeq[37]	RPKM (TMM,Upperquartile)	Non parametric	Null condition computed as contrast of fold changes and absolute difference within condition	✓	✓	×
SAMSeq [38]	SamSeq Method	Non parametric	Wilcoxon Rank + resampling	✓	×	×

*NB = Negative Binomial, GLM= General Linear Model

2.3. Tools for Differential Expression Analysis

With the increase of popularity of RNA-Seq studies, a lot of statistic tools for the detection of differential expression of genes and transcripts have been developed. The tools that are shown in Table 3 differ in normalization method, the statistic assumptions of the count distribution, and the statistic test to detect the DE genes. The details of each method can be found in the publications of the presented methods.

Numerous papers tried to compare the DE tools using both real and simulated data sets. However, the results are not completely in agreement with the performance of the tools associated with the properties of the dataset (e.g. number of samples, replicates, and heterogeneity of the dataset). There is no particular method that performs better when compared with each other, as some methods have certain strengths when used with definite datasets [39], [40]. On the other hand, the non-parametric methods NOISeq and SAMSeq showed opposite performance. NOISeq performs well when the two conditions in the dataset have different dispersion [39]. While it has a good control of the false discovery rate, it becomes too conservative with higher number of replicates [31]. On the contrary, SAMSeq performs well in terms of precision even as it needs more replicates to achieve a good power of detection [31], [39]. Above all, it is strictly dependent on the data in terms of performances [31]. In almost all the studies, the performances of DESeq, edgeR and BaySeq resulted to be similar in terms of accuracy, control of the number of false positives and the sensitivity [40]-[42]. Limma and DESeq showed to work well even with small sample sizes, whereby they

showed a low rate of false positives and decent power of detection. DESeq is found to be the most conservative while Limma performs well under different conditions maintaining consistency of the results and EdgeR is less conservative with a higher power of detection, but performs less consistent under different conditions [31], [39]. BaySeq showed good performances in different cases but is strongly dependent on the dataset structure [31], [39]. Cuffdiff2 was found to have a high precision with a significantly low number of false positives; however, it has a lower power of detection at gene level especially with a higher number of replicates [31], [40]. However, one of the main advantages of Cuffdiff2 is the possibility to compute expression changes at gene and at transcripts level, addressing both the estimation of expression values at gene and isoform level and considering the variability across replicates in the same pipeline. DESeq2 resulted to be more powerful but less precise (higher number of false positive) when compared to DESeq [31].

In our opinion, there are couples of parameters when considering the choice of appropriate tool. Firstly, perform multifactor analysis in case of a complex experimental design. Secondly, perform a DE analysis with no replicates. Even though the presence of at least three biological replicates (in dataset with minimal genetic end environmental variation) is of vital importance for a DE experiment using RNA-Seq data [13], this is not always possible due to the experimental limit. In such cases edgeR, DESeq, Noiseq and Cuffdiff2 allow overcoming this issue. Thirdly, some of the tools (e.g. edgeR , cuffdiff2, NOISeq, and baySeq) allow to choose different normalization methods. In our opinion, the possibility to choose between different methods in a fast and easy way (i.e. same tool/pipeline) is an essential plus. Fourthly, the presence of a well documented manual with clear description of the statistical assumption, filter, and data manipulation is also a very important entity in considering an ideal tool [31]. This can be seen in the manuals of DESeq, edgeR and Limma with practical examples provided for end-user. Finally, the possibility to check for normalization problems and other confounding factors could support good statistics analysis. NOISeq performs a quality control of the data to test the normalization outcome which could affect the statistical analysis with an easy interpretable and visual output.

3. Systems Genetics

Unraveling the genetic background of complex traits and diseases is a complicated task because of its multifactorial nature. To date, a huge number of Genome-Wide Association studies (GWAS) have been performed to detect associations between genetic variants and diseases, but their main limitation is a failure to explain a large proportion of the total genetic variation present in complex traits and diseases. Besides, several micro-array gene expression studies have investigated the transcriptome and explained more of the biological background by pointing towards associated pathways and the inclusion of genetic interactions. During the last few years, technologies to study the different omics scales (e.g. genomics and transcriptomics) have improved dramatically, resulting in the generation of huge amounts of NGS data. As introduced, “systems genetics” integrates omics scale measurements in a holistic analyses model to unravel the genetic and corresponding biological background of complex traits and diseases [6]. This means that all the different biological levels are to be integrated including the phenome, genome, epi-genome, transcriptome, proteome and metabolome to fully understand the disease or trait. Current technologies provide us the opportunity to get a highly accurate measurement of each layer. As a result, a more complete picture of the biological and genetic pathways leading to understanding particular complex diseases is obtained.

3.1. Pathway Analyses

Pathway analyses are originating from the micro-array gene expression analysis [43], whereby prior biological knowledge is used to identify the functional annotation of groups of genes and corresponding

pathways with the goal to gain more biological insight into the disease/trait under the study. There are some differences between micro-array expression data and RNA-Seq expression data (e.g. normal distribution *vs*. NB distribution)with specific RNA-Seq tools for pathway analysis.

One of the biases in RNA-Seq data is the gene length bias, which means that longer genes have a higher chance of being sequenced more often thereby lead to higher counts. Consequently, those genes have also a higher change of being detected as "differentially expressed" due to a higher statistical power [44]. GO-Seq is an R-Bioconductor package that performs gene ontology analysis on differentially expressed genes including a correction for the gene length bias [45]. GO-Seq inputs a list of DE genes and the complete set of genes (background), incorporates the transcript length of the DE genes into the statistical tests and gives the significance level of KEGG pathways or GO terms present as output. Another recently published method for pathway analysis using RNA-Seq data is GSAASeqSP/GSAASeqGP [46], which implements both sample/phenotype permutation and gene permutation (recommended with sample sizes above seven per phenotype) and GSAASeqGP only gene permutation. GSAASeq is a part of the GSAA software suit that is freely available for non-commercial use and can be downloaded upon registration. Whereas GSAASeq requires RNA-Seq data from two distinct phenotypic groups (e.g. case-control), it detects differentially expressed genes using Signal2Noise_log2Ratio and further calculates gene set statistics using WeightedSigRatio or SigRatio. Other options within the GSAA software suit have not been evaluated yet for RNA-Seq datasets. Besides those RNA-Seq specific pathway analysis tools, there are many other publically available tools available like DAVID [47], GOEAST [48] and GSEA [49]. In addition, several commercial software suits offer pathway analysis based on RNA-Seq specificity.

3.2. Network Analyses

It is well known that genetic interactions play a major role in biology and it is expected that by investigating them, we may reveal important knowledge about the genetic and biological architecture of diseases. One of the ways to study such interactions is using a network analysis approach. Network approaches focus on the clustering of genes instead of the individual genetic variant, with the main credence that genes in those clusters are someway functionally related to each other [50]. Consequently, gene networks might provide a better understanding of the pathways where genes are ranked, for example, based on their regulating function in a cellular event. Gene networks are often graphically represented with genes presented as nodes and their association as edges, with edges directed or undirected. The degree (also called "connectivity") of a particular node is the sum of connections or connection strengths with other nodes and the distribution of this degree represents the probability that a particular node has a certain number of connected nodes. In biology, networks are mostly scale-free, meaning that the degree distribution follows a power-law [51]. Further a scale-free biological network consists of a very few genes with a high degree (also called hubgenes) and a very large number of genes with a low degree. Hub genes are biologically very essential and even important as it has been shown that a problem in a hub gene causes breakdown of the network [52].

There are several approaches available to build gene networks and study the interactions between genes, *viz.* Co-expression/regulatory patterns [53], Bayesian networks [54] or Random Forest Tree approaches [55] or Artificial Neural Networks (ANN) [56]. Those different network methods might lead to different results, however, it has been shown that there is no specific method better than others and the integration of different network methods could provide a more complete picture of the interactions present in the dataset [57].

3.2.1. Co-expression network approaches

Gene co-expression networks calculate the interaction between pairs of genes based on the correlation

between the expression patterns of those pairs of genes. One of the well-known methods is Weighted Gene Co-expression Network Analysis (WGCNA) which calculates the Pearson's correlation between all genes based on the sum of correlations of a particular gene with all other genes [53], [58]. The clusters of genes are detected based on the Topological Overlap Measure (TOM), which represents the number of shared neighbors between a pair of genes as a value between 0 and 1, representing no shared neighbors and same neighbors between the two genes respectively. Based on the TOM, a gene dendrogram is created and clusters of genes (modules) are detected by cutting the branches of the dendrogram. The key drivers of the modules are based on the phenotype(s) of interest and intra-modular characteristics. Previously, we have applied this method successfully on RNA-Seq data in a pig model for human obesity [3].

3.2.2. Regulatory network approaches

There are a wide range of biological processes effectively used as regulatory molecules. Right from the PPI models used as regulatory models to the existing class of ncRNAs, there has been a great attention and focus on how regulatory networks modulate their role through interactions. Characterizing transcript structures and understanding expression profiles mediating regulatory roles have relatively come up with the ENCODE project [59]. However, very little on how the lncRNAs regulate interactions between noncoding RNA classes is known. Recent reports hypothesized how lncRNAs contribute towards regulatory interactions with its other non-coding peers like miRNAs [60]. This analysis advocates the use of such regulatory interactions between classes of ncRNAs classes and their functional implication.

3.2.3. Bayesian network approaches

Bayesian network (BN) approach is another way to model RNA-Seq expression data to find relationships among genes without previous biological assumptions. From a mathematical point of view, a BN is used to represent a joint probability distribution of random variables. Additionally, BN uses directed acyclic graph where the vertices are a set of random variables and their edges being conditional dependencies. Different methods are used during the learning, evaluation and inference phase such as Gibbs sampler that allows approximating from a specified multivariate probability distribution. One of the major advantages of the BN is that it allows distinguishing direct and undirected associations. This together can be used to integrate genetic data with gene expression data further allowing us to identify causality [5], [61]. In conclusion, the BN provides a deep insight into the biological mechanism of a biological process. Between BN and Lemon Tree, a Gibbs sample based algorithm is often used to find clusters of genes based on their co-expression values and the genes that play a regulatory role in each module [62]. The tool is a platform independent command line in Java and is a structured modular program that runs singularly where the output of the previous task is the input of the following one. Concisely, the first task in Lemon Tree runs instances of a model as Gibbs sampler infers co-expression modules and condition clusters within module. Further the different module structures obtained with different runs of the first task are used to build a consensus set of modules considering the frequency of each pair of nodes in the same cluster. Finally, a set of regulatory programs for a set of modules are computed with the computation of decision trees, the score is then assigned considering the number of trees in which the regulator is allocated, a significance level is then computed using an empirical distribution of scores.

3.2.4. Random forest tree approaches

Large-scale genomic data heralds a great challenge for statisticians and bioinformaticians owing to the high aspects of genomic features the data is compounded with. The approaches discussed above may have a highly correlated structured genomic data and less order of gene or protein interactions. Although a good number of learning methods could be used to understand how networks fit to understand the variables in

the interactions, they fall short of accurate data mining. Brieman L in 2001 proposed one such collective learning method called Random Forest Tree, which can be widely applied to predict the nearest neighbors [55]. Random Forest Tree algorithm ranks genes, which can be further used for unsupervised learning (/clustering), using large data sets which the data need not follow any specific distribution [63]. By this way, another approach is offered to gain knowledge in understanding expression data.

3.2.5. Artificial neural networks

Artificial Neural Networks (ANN) is a method of artificial intelligence based on human brain functionality [56]. ANNs are nonlinear pattern recognition techniques that can be used as a tool in medical decision making. An ANN basic design consists of three cycles, learning, testing and decision making. A learning strategy is applied to change the weights in order to optimize the error. ANNs recognizes patterns in the data from a known dataset called the training data and the main goal of the network is to make predictions on novel inputs called the test data. During learning cycle, a function is optimized to maximize the capture of positives and rejection the negative data points. In the iterations over a number of cycles called "epochs", every data point in training data is fed to the ANN one after the other. The error in prediction is calculated and weights are updated [64]. A particular type of ANN called Recurrent Neural Network (RNN) has been developed to find out gene regulatory networks in time series RNA expression data [65], [66], wherein positive and negative feedback loops are considered on the internal states. The RNNs have significant characteristics to make it computationally feasible (e.g. resistance to noise and non-linearity [67]) in analyzing RNA-Seq data in combination with other clustering approaches [68].

3.3. Data Integration

As mentioned, systems genetics approaches are based on the integration of different 'omics data levels. One such way is the integration of genomics and transcriptomics by detecting expression QTLs (eQTLs). An eQTL is a genomic region associated with transcript levels, which subsequently affects the phenotype. It has been shown that eQTLs are highly heritable [69] and they might provide more information on the biological control of gene expression [70], but also provide more knowledge about the function of a genetic variant for example detected using a GWAS. The eQTLs can be cis or trans-acting, in the case of a cis-eQTL, the eQTL is near the location of the gene encoding the transcript, *i.e.* closed to the transcription start site (TSS) while in case of a trans-eQTL the eQTL is on a larger distance or even on another chromosome of the gene encoding the transcript. Cis-eQTLs have generally large effect sizes, but their effect size generally increases when the distance between the eSNP and the TSS increases. The exact working mechanism behind the trans-eQTL is not known, but often effect sizes are small. However, studies have shown that trans-eQTLs provide valuable insight into disease pathogenesis. Using the same approach as with eQTL mapping, we can also integrate the metabolome with the genome: metabolomic QTL (mQTL) mapping [71]. The metabolome includes all measurable metabolites in a cell, such as lipids, carbohydrates and amino acids. It has been shown that genetic variants (detected by e.g. GWAS) have a major effect on the metabolome [72] and therefore mQTL mapping might help identify genetic variants affecting the phenome on the metabolome.

Another way of integrating different omics data levels is using public available datasets. Here, we can think of the well-known pathway analysis using databases of gene ontology and pathways to indicate the biological background of findings while including a different 'omics data levels such as PPI in predicting the function of associated genes. While STRING, a database with known and predicted protein associations and interactions is used to automatically mine and find association of genes [73], GeneMANIA, a cytoscape web-based tool is used to interpret the network weights based on size of input gene list [74]. This in-turn is based on Gene-Ontology (GO) based weighting, which assumes the set of genes have a biological processes as defined by GO.

In the recent, long noncoding RNAs (lncRNAs) constituting transcriptomic data have been employed to ascertain their functional relationship. Studies on functional characterization of lncRNAs have resulted in interaction data with their RNA peers, DNA or proteins [75]. There have been no significant pipelines to address the coexpression networks regulating the coding-noncoding genes. However, by and large, annotated lncRNA data could be used to construct a co-expression network based on existing gene expression profiles specific to various diseases, *viz.* cancers, diabetes and autoimmune disorders. Zhao *et al*. have recently established such functional annotation pipeline allowing the researchers to predict the function of lncRNAs [76]. Although the methods employ hub-based and model-based networks as discussed above, it would be interesting to see how the data could be used to predict the better association of lncRNAs with existing coding genes. Whether or not other models such as BN could appropriately be used remains to be understood. Keeping in view of the larger diversity of function of lncRNAs, explosive growth in functional relationships of lncRNAs specific to diseases can be categorically approached. From the aforementioned network approaches, it would be interesting to see a comparative functional genomics approach to distinguish non-coding regions from protein coding regions. Finding potential archetypes for lncRNAs that are involved in disease genotypes modulating genotype-phenotype data could be interesting as a part of RNA-seq studies. It would be remarkable to see if RNA-Seq data has some questions answered on evolutionary existence of ncRNAs. If the data has lncRNAs, we can explore the possible role or lncRNAs in diseases thus enabling lncRNAs serve as a metaphor to RNA-Seq data.

4. Applications of Systems Genetics in Animals

As RNA-Sequencing is getting more and more popular with decreased costs and time, the number of animal-based studies using RNA-Seq data are increasing. We recently published our first study using RNA-Seq data in a porcine model for human obesity using a network approach [77]. The porcine model used was an F2-population, created by crossing the Göttingen Minipig (prone to obesity) with Duroc and Yorkshire sows (bred for centuries for lean meat content). The resulting ~500 F2-pigs were all deeply phenotyped and genotyped using the Illumina 60K Porcine SNP-chip. The degree of obesity of each individual pig was represented by the Obesity Index (OI), an aggregate genotypic value, and used for selection of 36 animals from the resource population for RNA-Sequencing of subcutaneous adipose tissue. A gene co-expression network was constructed using the WGCNA approach, whereby we detected in total 20 modules of highly interconnected genes. Based on their association with obesity-related phenotypes, we selected five modules for functional annotation using GO-Seq to correct for gene length bias. The biologically most interesting module was the Blue Module, consisting of 69 genes, showing an over-expression in the obese individuals. Functional annotation showed a strong association for immune-related pathways (e.g. *Natural Killer Cell Mediated Cytotoxity* and *B cell receptor signaling pathway*), but interestingly, the strongest association was found for "*Osteoclast differentiation*" (P-value = $1.4E^{-7}$). To find potential causal genes in this module, we used a Bayesian approach using the Lemon-Tree software suit. Lemon-Tree detected three regulator genes: *CCR1*, *MSR1* and *SPI1.* Interestingly, those three genes have all been individually associated already with bone remodeling, e.g. in mice studies. Based on our results, we suggest a potential causal role for those genes in the association between obesity and osteoporosis possibly via the immune system. We inferred systems genetics approaches in detecting novel genes and pathways better insight in understanding complex traits and diseases.

5. Conclusions

In this review, we have shown the potential of using RNA-Seq data in systems genetics approaches to study complex traits and diseases. The possibility to obtain a huge amount of transcriptomics and genomic

data became faster and cheaper, thanks to the expansion of NGS techniques. This huge amount of data has a great potential in studying the causal genetic and other biological processes through the integration of 'omics data coming from multiple biological levels.

Several studies have shown that RNA-Seq data from high throughput techniques needs a prerequisite quality control of the aligned data, and the choice of the analytical methods could have a great impact on the reliability of the final results. We have presented several post-mapping QC tools, which result in different type of statistics. In order to generate a complete report of the quality of the reads, we propose that the output of multiple tools should be merged together before proceeding to the analysis. For instance, we have discussed how QUACRS allows to do such tasks in an easy interpretable way for users with less informatics skills.The normalization can also have a strong impact on the final outcome, influencing the results of both the DE analysis and any other systems genetics approach. The choice, however, could be based on the final objective, as some of the normalization methods are necessary when aimed to rank the genes "within samples", while others are more effective when a "between samples" comparison is performed. In the last case, the TMM and DESeq scaling factor was found to have best performances allowing accurate sequencing depth when considering the highly expressed genes in the dataset.

Continuing with the analysis of RNA-Seq data, we further reviewed several tools for DE analysis and systems genetics analyses, e.g. network methods. In particular the outcomes and the performances of the DE tools seemed to be dependent on the dataset structure and characteristics. Some of the tools showed a consistency in the precision of the results across different datasets (Limma and DESeq), but the possibility to test and compare different tools with the real dataset under question still seems to be the best option. In general, system genetics approaches give the opportunity to get a deeper insight into the biological and genetic architecture of complex traits and diseases. Using network methods and pathway analysis, we gain insight in biological processes, but also due to data integration, we get a better overview of the biological systems leading to diseases. In principle, we expect to have a paradigm shift of RNA-Seq data with the impending list of ncRNAs (in specific, lncRNAs) integrating genomics and transcriptomics.

Acknowledgment

Gianluca Mazzoni is funded by the PhD grant within the GIFT project (http://gift.ku.dk/) from Programme Commission on Health, Food and Welfare of the Danish Council for Strategic Research (Innovationsfonden). Lisette J. A. Kogelman is funded by post-doctoral fellowship within Biochild project (http://biochild.ku.dk/) the Programme Commission on Individuals, Disease and Society under the Danish Council for Strategic Research (Innovationsfonden). Prashanth Suravajhala and Haja N. Kadarmideen thank EU-FP7 Marie Curie Actions – Career Integration Grant (CIG-293511) for partially funding their time spent on this research.

References

[1] Kadarmideen, H., Von Rohr, P., & Janss, L. (2006). From genetical genomics to systems genetics: Potential applications in quantitative genomics and animal breeding. *Mammalian Genome*, *17(6)*, 548-564.

[2] Li, H. (2013). Systems genetics in "-omics" era: Current and future development. *Theory Biosci., 132(1)*, 1-16.

[3] Kogelman, L. J. A., *et al*. (2014). Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA sequencing in a porcine model. *BMC Medical Genomics*, *7*, 57.

[4] Kadarmideen, H. N., & Janss, L. L. (2007). Population and systems genetics analyses of cortisol in pigs

divergently selected for stress. *Physiol Genomics*, *29(1)*, 57-65.

[5] Civelek, M., & Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, *15(1)*, 34-48.

[6] Kadarmideen, H. N. (2014). Genomics to systems biology in animal and veterinary sciences: Progress, lessons and opportunities. *Livestock Science*, *166*, 232-248.

[7] Fu, J., *et al*. (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet*, *8(1)*, e1002431.

[8] Westra, H. J., & Franke, L. (2014). From genome to function by studying eQTLs. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1842(10)*, 1896-1902.

[9] Cooper, G. M., & Shendure, J. (2011). Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*, *12(9)*, 628-640.

[10] Zhang, B., *et al*. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*, *153(3)*, 707-720.

[11] Matsui, M., *et al*. (2013). Promoter RNA links transcriptional regulation of inflammatory pathway genes. *Nucleic Acids Res.*, *41(22)*, 10086-10109.

[12] Shen, K., & Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, *26(10)*, 1316-1323.

[13] Williams, A. G., *et al*. (2014). RNA-seq data: Challenges in and recommendations for experimental design and analysis. *Curr Protoc Hum Genet*, *83*, 1-20.

[14] Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. From http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[15] Wysoker, A., Tibbetts, K., & Fennell, T. (2012). Picard. From http://picard.sourceforge.net/

[16] Garcia-Alcalde, F., *et al*. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*, *28(20)*, 2678-2679.

[17] DeLuca, D. S., *et al*. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, *28(11)*, 1530-1532.

[18] Wang, L., Wang, S., & Li, W. (2012). RSeQC: Quality control of RNA-seq experiments. *Bioinformatics*, *28(16)*, 2184-2185.

[19] Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2011). SAMStat: Monitoring biases in next generation sequencing data. *Bioinformatics*, *27(1)*, 130-131.

[20] Kroll, K. W., *et al*. (2014). Quality control for RNA-seq (QuaCRS): An integrated quality control pipeline. *Cancer Informatics*, *13(Suppl 3)*, 7.

[21] Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*, *4(1)*, 14.

[22] Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biol*, *11(12)*, 220.

[23] Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, *11(10)*, R106.

[24] Yang, Y. H., & Thorne, N. P. (2003). Normalization for two-color cDNA microarray data. *Lecture Notes-Monograph Series*, 403-418.

[25] Bolstad, B. M., *et al*. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19(2)*, 185-193.

[26] Mortazavi, A., *et al*. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, *5(7)*, 621-628.

[27] Wang, K., *et al*. (2011). A genome-wide association study on obesity and obesity-related traits. *PLoS*

One, *6(4)*, e18939.

[28] Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, *11(3)*, R25.

[29] Bullard, J. H., *et al*. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11(1)*, 94.

[30] Dillies, M. A., *et al*. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, *14(6)*, 671-683.

[31] Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform*, *16(1)*, 59-70.

[32] Hardcastle, T. J., & Kelly, K. A. (2010). Bayseq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, *11(1)*, 422.

[33] Trapnell, C., *et al*. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31(1)*, 46-53.

[34] Love, M., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15(12)*, 550.

[35] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26(1)*, 139-140.

[36] Ritchie, M. E., *et al*. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, *43(7)*, e47.

[37] Tarazona, S., *et al*. (2012). NOIseq: A RNA-seq differential expression method robust for sequencing depth biases. *EMBnet. Journal*, *17(B)*, 18-19.

[38] Li, J., & Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, *22(5)*, 519-536.

[39] Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, *14*, 91.

[40] Zhang, Z. H., *et al*. (2014). A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS One*, *9(8)*, e103207.

[41] Rapaport, F., *et al*. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*, *14(9)*, R95.

[42] Kvam, V. M., Liu, P., & Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.*, *99(2)*, 248-256.

[43] Emmert-Streib, F., & Glazko, G. V. (2011). Pathway analysis of expression data: Deciphering functional building blocks of complex diseases. *PLoS Computational Biology*, *7(5)*, e1002053.

[44] Oshlack, A., & Wakefield, M. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, *4(1)*, 14.

[45] Young, M., *et al*. (2010). Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biology, 11(2)*, R14.

[46] Xiong, Q., Mukherjee, S., & Furey, T. S. (2014). GSAASeqSP: A toolset for gene set association analysis of RNA-seq data. *Sci. Rep.*, *4*, 6347.

[47] Huang, D., Sherman, B., & Lempicki, R. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, *4*, 44-57.

[48] Zheng, Q., & Wang, X. J. (2008). GOEAST: A web-based software toolkit for gene ontology enrichment analysis. *Nucleic Acids Research*, *36(suppl 2)*, W358-W363.

[49] Subramanian, A., *et al*. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci., USA*, *102*, 15545-15550.

[50] Brazhnik, P., de la Fuente, A., & Mendes, P. (2002). Gene networks: How to put the function in genomics. *Trends in Biotechnology*, *20(11)*, 467-472.

[51] Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, *118(21)*, 4947-4957.

[52] Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74(1)*, 47-97.

[53] Horvath, S. (2011). *Weighted Network Analysis — Applications in Genomics and Systems Biology*. Verlag, New York: Springer.

[54] Friedman, N. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, *7*, 601-620.

[55] Breiman, L. (2001). Random forests. *Machine Learning*, *45(1)*, 5-32.

[56] McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5(4)*, 115-133.

[57] Marbach, D., *et al*. (2012). Wisdom of crowds for robust gene network inference. *Nat. Meth*., *9(8)*, 796-804.

[58] Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9(1)*, 559.

[59] Consortium, E. P. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science*, *306(5696)*, 636-640.

[60] Jalali, S., *et al*. (2013). Systematic transcriptome wide analysis of lncRNA-miRNA interactions. *PLoS One*, *8(2)*, e53823.

[61] Nagrecha, S., Lingras, P. J., & Chawla, N. V. (2013). Comparison of gene co-expression networks and bayesian networks. *Proceedings of the 5th Asian Conference on Intelligent Information and Database Systems - Volume Part I* (pp. 507-516). Kuala Lumpur, Malaysia: Springer-Verlag.

[62] Bonnet, E., Calzone, L., & Michoel, T. (2015). Integrative multi-omics module network inference with Lemon-Tree. *Preprint ArXiv*, *11(2)*, e1003983.

[63] Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99(6)*, 323-329.

[64] Yadav, R. (2015). Translating inter-individual genetic variation to biological function in complex phenotypes. PhD Thesis, Danish Technical University, Denmark.

[65] Vohradský, J. (2001). Neural network model of gene expression. *The FASEB Journal*, *15(3)*, 846-854.

[66] Noman, N., Palafox, L., & Iba, H. (2013). Reconstruction of gene regulatory networks from gene expression data using decoupled recurrent neural network model. In Y. Suzuki, & T. Nakagaki (Eds.), *Natural Computing and Beyond* (pp. 93-103), Japan: Springer.

[67] Lee, W. P., & Tzou, W. S. (2009). Computational methods for discovering gene networks from expression data. *Briefings in Bioinformatics*, *10(4)*, 408-423.

[68] Lee, W. P., & Yang, K. C. (2008). A clustering-based approach for inferring recurrent neural networks as gene regulatory networks. *Neurocomputing*, *71(4-6)*, 600-610.

[69] Dixon, A. L., *et al*. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, *39(10)*, 1202-1207.

[70] Cookson, W., *et al*. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, *10(3)*, 184-194.

[71] Dumas, M. E., & Gauguier, D. (2012). Mapping metabolomic Quantitative Trait Loci (mQTL): A link between metabolome-wide association studies and systems biology. In K. Suhre (Ed.), *Genetics Meets Metabolomics* (pp. 233-254), New York: Springer.

[72] Gieger, C., *et al*. (2008). Genetics meets metabolomics: A Genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, *4(11)*, e1000282.

[73] Franceschini, A., *et al*. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*., *41(Database issue)*, 808-815.
[74] Montojo, K., *et al*. (2010). Gene MANIA cytoscape plugin: Fast gene function predictions on the desktop. *Bioinformatics*, *26(22)*, 2927-2928.
[75] Novikova, I. V., *et al*. (2013). Rise of the RNA machines: Exploring the structure of long non-coding RNAs. *Journal of Molecular Biology*, *425(19)*, 3731-3746.
[76] Zhao, Y., *et al*. (2014). Computational methods to predict long noncoding RNA functions based on co-expression network. *Methods Mol*. *Biol*., *1182*, 209-218.
[77] Kogelman, L. J. A., *et al*. (2014). Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA sequencing in a porcine model. *BMC Medical Genomics*, *7(1)*, 57.

Gianluca Mazzoni was born in Ancona, Italy on May 1, 1989. In 2011, he finished his bachelor in biotechnology at Bologna University, Italy. In 2013, he finished his master in bioinformatics at Bologna University, Italy. In 2014 he started his PhD at the Department of Veterinary, Clinical and Animal Sciences at the University of Copenhagen, Quantitative Genetics, Genomics and Systems Biology Group, where he is currently employed.

Lisette J. A. Kogelman was born in Raalte, The Netherlands on May 6, 1987. In 2009 she finished her bachelor in animal sciences at Wageningen University and Research Centre, The Netherlands. Then, in 2011, she finished her master in animal sciences with a specialization in adaptation physiology and animal genetics, likewise at Wageningen University and Research Centre, The Netherlands. In 2011, she started her PhD at the Department of Veterinary, Clinical and Animal Sciences at the University of Copenhagen, Denmark. Her work was performed in the quantitative genetics, genomics and systems biology group under the supervision of prof. Haja N. Kadarmideen. Currently she is working as a postdoctoral researcher in this group, focusing on systems genetics approaches to analyse different omics levels.

Prashanth Suravajhala was born in Kothagudem, India on January 1, 1979. He finished his master in biotechnology from Guru Ghasidas Central University, Bilaspur, India in 2002. He worked as a graduate research fellow for Centre for Cellular and Molecular Biology (CCMB), Hyderabad, India before taking up his PhD position. He obtained the PhD in soft and biological matter from Aalborg University in 2012. He went on to do his postdoctoral sabbatical working for short stints at protein data bank of Japan and Bioinformatics.org. He has 10 or more years' experience in functional genomics and bioinformatics and protein-protein interactions. He has published 17 peer-reviewed journal publications and authored three books. He is currently managing bioinformatics resources at quantitative systems genetics group, Department of Veterinary Clinical and Animal Sciences, University of Copenhagen since January 2015. His interests are functional genomics of hypothetical proteins and regulatory regions especially those that include lncRNAs and miRNAs. He has been a member of International Society for Computational Biology (ISCB), American Association for Advancement in Science (AAAS) since 2005. He is an associate director of Bioinformatics.Org, which advocates open access. In 2005, he founded a virtual organization called

Bioclues.org, which masks mentor-mentee relationships and helps bioinformatics students move forward with their career goals. He received Bayer visiting postdoc fellowship in 2014 apart from few visiting grants/fellowship towards his postdoctoral visits.

Haja Kadarmideen is an Australian citizen. He was born in India in 1967. Haja Kadarmideen is currently a full professor and group leader of animal breeding, quantitative genetics and systems biology in the Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark. He obtained a doctor of veterinary medicine (DVM) and master degree (MVSc) in veterinary genetics (Chennai, India) in 1989 and 1992, respectively. He obtained his PhD degree in quantitative genetics from the University of Guelph (Ontario, Canada) in 1998. He was a scientist in dairy cattle genetics for over 3 years at the Scottish Agricultural College, Edinburgh, UK. He was then the head of statistical genetics group at the Swiss Federal Institute of Technology (ETH) Zurich in Switzerland (2001-2006). In 2006, he was appointed as a principal scientist and the leader of Quantitative Genetics and Systems Biology Group at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia, where he worked until his moved to the University of Copenhagen in January 2011. In addition to his full professorship, he is also the director of 3 major national/international consortiums and serves as a board and steering committee member of various research and innovation initiatives nationally. He has authored/co-authored over 230 scientific publications (journal articles with scientific peer review, refereed book chapters, conference proceedings papers, abstracts etc). He is on the editorial board of *Frontiers in Genetics*, *Frontiers in Veterinary Science*, *Turkish J. of Veterinary & Animal Sciences* and *EC Veterinary Science* journals. His group's website is www.qsg.dk.